# CSCI 6333 Data Mining & Warehousing

**Module 5: Anomaly Detection, Avoiding False Discoveries**

**Homework Assignment Four**

**All problems are equal-weighted with 20 points each.**

1. Consider distance-based outliers with respect to two parameters and : A data point is an outlier if the number of points within a radius to is less than . Explain in detail one major weakness of this formulation.

**Answer 1.** The major weakness of the formulation is that it doesn’t consider the probable density of the data points. For example, comparing the clusters with the same number of points, the cluster with less density is more vulnerable for the formulation above. To avoid this, either the density-based or the nearest-neighbor based approach might be used.

1. Consider the density-based outliers. Describe in detail the time complexity of finding local outlier factors (LOFs) for all data points in a set of points. Assume that the dimension is and the parameter k is given.

**Answer 2.** Firstly, the k-nearest neighbors of all objects should be calculated. Then, we calculate the number of k-neighbors for each data point. The complexity is either O(N\*d\*k) or O(N\*(d+k)) time, which depends on the computation strategy. Then, for each data pair (A, B), we calculate the Reachability Density. It takes O(N2\*d) time. Then, the Local Reachability Density for each data point is calculated. This also takes O(N2\*d) time (the number of pairs in the previous step). In the final, the Local Outlier Factor is calculated. Since each data point has only one LRD, it takes O(N\*d) time. So the total time complexity is **[O(N\*d\*k) or O(N\*(d+k))] + O(N2\*d) + O(N2\*d) + O(N\*d) = [O(N\*d\*k) or O(N\*(d+k))] + O(N2\*d)**. For very small k and d, we can write it as **O(N2)**

1. Consider the connectivity-based outliers. Describe in detail the time complexity of finding connectivity outlier factors (COFs) for all data points in a set of points. Assume that the dimension is and the parameter k is given.

**Answer 3.** Firstly, the k-nearest neighbors of all objects should be calculated. Then, we calculate the number of k-neighbors for each data point. The complexity is either O(N\*d\*k) or O(N\*(d+k)) time, which depends on the computation strategy……..

1. Explain in detail two different methods to use k-means clustering to detect outliers.

**Answer 4.** There are two different methods of clustering: Finding anomalous clusters that are distant from the other clusters and finding anomalous instances of clusters. By using k-means clustering, we can use both of the methods.

The method of finding anomalous clusters:

1) Do K-means clustering;

2) Get the centroids of clusters;

3) Calculate the distance between centroids;

4) If all distance’s of centroid is significantly higher than the average distance; between centroids (for example, for 3 clusters, 1st is far from both 2nd and 3rd, define the centroid and all its points as outliers.

The method of finding anomalous instances:

1) Do K-means clustering;

2) Calculate the distance between centroids and its points;

3) Calculate the average distance between centroids and its points;

4) If all distance’s of point is significantly higher than the average distance between point and its centroid, define the point as an outlier.

1. Consider the problem of determining whether a coin is a fair one, i.e. P(heads)=P(tails)=0.5, by flipping the coin 10 times. Use the binomial theorem and the basic probability to answer the following questions.
   1. A coin is flipped ten times and it comes up heads every time. What is the probability of getting 10 heads in a row and what would you conclude about whether the coin is fair?
   2. Suppose 10,000 coins are each flipped 10 times in a row and the flips of 10 coins results in all heads, can you confidently say that these coins are not fair?

**Answer 5.**

a) The probability to get head 10 times in the row is equal to: P(10) = (10, 10)\*(0.5)10\*(1-0.5)(10-10) = **0.001**

Our null hypothesis: The coin has both the head and the tail, so the coin is fair.

The mean of distribution is equal to μ = 0.009765625\*1 + 0.0439453125\*2 + 0.1171875\*3 + 0.205078125\*4 + 0.24609375\*5 + 0.205078125\*6 + 0.1171875\*7 + 0.0439453125\*8 + 0.009765625\*9 + 0.0009765625\*10 = 5

The variance of distribution is equal to σ2 = (0.009765625\*12 + 0.0439453125\*22 + 0.1171875\*32 + 0.205078125\*42 + 0.24609375\*52 + 0.205078125\*62 + 0.1171875\*72 + 0.0439453125\*82 + 0.009765625\*92 + 0.0009765625\*102) – μ2 = 27.5-52 = 2.5, so σ = sqrt(2.5) ≈ 1.5811

For our case x=10, Z-score is equal to ((5-μ)/σ)=0

By both one-right-tailed test, for α=0.05, we can be confident for 95% that the null hypothesis is wrong.

Our alternative hypothesis: The coin is 99.99% biased.

The probability to get head 10 times in the biased coin is equal to: P(10) = (10, 10)\*(0.99)10\*(1-0.99)(10-10) = **0.9044**

The mean of distribution is equal to μ = 9.90000000000008E-18\*1 + 4.41045000000003E-15\*2 + 1.16435880000001E-12\*3 + 2.01725162100001E-10\*4 + 2.39649492574801E-08\*5 + 1.97710831374211E-06\*6 + 0.000111847841749\*7 + 0.004152351124926\*8 + 0.091351724748364\*9 + 0.904382075008804\*10 = 9.9

The variance of distribution is equal to σ2 = (9.90000000000008E-18\*12 + 4.41045000000003E-15\*22 + 1.16435880000001E-12\*32 + 2.01725162100001E-10\*42 + 2.39649492574801E-08\*52 + 1.97710831374211E-06\*62 + 0.000111847841749\*72 + 0.004152351124926\*82 + 0.091351724748364\*92 + 0.904382075008804\*102) – μ2 = 98.109-9.92 = 2.5, so σ = sqrt(0.099) ≈ 0.3146

For our case x=10, Z-score is equal to ((10-μ)/σ)=0.3178

Since the Z-score of our hypothesis is higher than the Z-score of the null hypothesis, we can reject the null hypothesis and prove that our hypothesis is correct.

b)